



## DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

### Analítica Computacional para la toma de Decisiones – 202420

#### Proyecto 1 - Bicicletas compartidas

**Integrantes: Jerónimo Vargas – 202113305, Carlos Gómez – 202111593,  
Juan Esteban López – 202021417**

#### Asignación de Roles

- **Ingeniería de datos** – Jerónimo Vargas
- **Análisis de datos** – Carlos Gómez
- **Ciencia de datos** – Carlos Gómez
- **Análisis de negocio** – Juan Esteban López
- **Tablero de datos** – Juan Esteban López
- **Despliegue** – Jerónimo Vargas

#### Preguntas de negocio y plan de acción

##### Tarea 1

#### Enfoque del proyecto

Se escogió el segundo enfoque de cliente el cual está más inclinado hacia mejorar la rentabilidad del negocio. Para poder realizar el mejor producto posible para el cliente queremos responder las siguientes preguntas:

1. ¿Cuál es el costo que debe tener una bicicleta por hora para que el proyecto sea viable teniendo en cuenta los costos y la demanda?
2. ¿Cuál es el costo del sistema y como se distribuyen en costos de operación y costos de mantenimientos?
3. ¿Cuál sería el pronóstico de demanda de bicicletas en cada hora del día dadas unas condiciones?

Mediante visualizaciones de datos se puede mostrar la distribución de los costos en el sistema. Además, se puede realizar gráficos en los que se puedan ver patrones en la demanda los cuales pueden explicar el pronóstico de demanda que se haga para el futuro.

Con un modelo de regresión lineal se espera dar el mejor pronóstico posible de la demanda de bicicletas para cada hora del día con las variables proporcionadas. Se realizarán pruebas estadísticas para determinar cuáles de las variables son útiles para el modelo y así poder otorgar un modelo predictivo lo más acertado posible. Con un buen pronóstico de la demanda y guiados por modelos de negocio similares en diferentes países se podrá estimar la tarifa mínima en las condiciones dadas para cubrir todos los costos.

Las preguntas planteadas anteriormente se responderán a lo largo del proyecto. Sin embargo, a grandes rasgos se otorgará un precio sugerido para que la empresa pueda cubrir sus costos y pueda llegar a una rentabilidad esperada en una hora del día en específico con condiciones dadas. Además, se pedirán por parámetro los costos de la empresa para poder hacer un análisis con la demanda pronosticada y poder determinar cómo se distribuyen los costos del sistema. Finalmente se brindará un pronóstico de la demanda en una hora del día en específico con las condiciones que quiera simular el usuario y así poder planear la operación.

## Datos

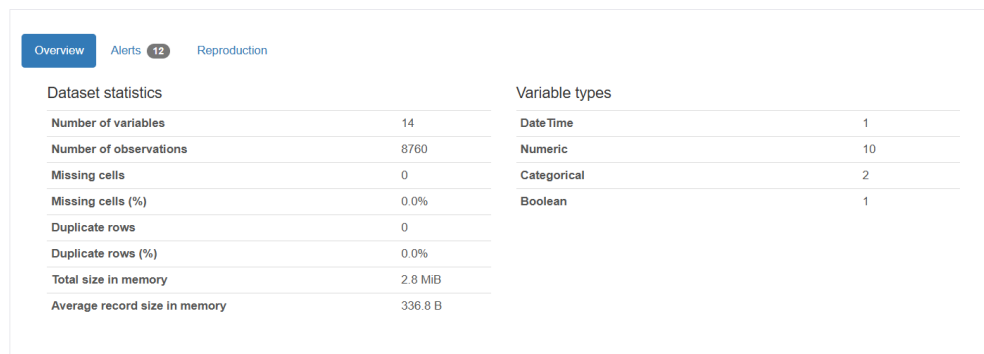
### Tarea 2 - Limpieza y alistamiento de datos

Carga y limpieza de datos en Python, identificación y gestión de datos faltantes, tratamiento adecuado de variables categóricas para análisis y modelamiento posterior, asegurando que los datos estén en un formato óptimo.

Todo el procedimiento realizado en esta anterior lo pueden encontrar en Jupyter Notebook “DataAnalysisClean.ipynb”.

Primero, para realizar la exploración de los datos, se utilizó una librería llamada ydata-profiling (anteriormente conocida como pandas-profiling). Esta herramienta permite realizar un perfilamiento completo de los datos, proporcionando un análisis detallado que facilita la identificación de problemas como valores nulos, duplicados, tipos de datos incorrectos, y la distribución de las variables. A través de este perfilamiento, se puede evaluar la calidad del dataset y determinar qué pasos de limpieza y preprocesamiento son necesarios antes del análisis o modelamiento posterior.

En la imagen proporcionada, se observa el resumen del perfil de los datos generados por ydata-profiling. Los principales hallazgos del reporte incluyen:



The image shows a screenshot of the ydata-profiling report overview. It has three tabs: Overview (selected), Alerts (12), and Reproduction. The Overview tab displays two tables: 'Dataset statistics' and 'Variable types'.

Dataset statistics	
Number of variables	14
Number of observations	8760
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	2.8 MiB
Average record size in memory	336.8 B

Variable types	
DateTime	1
Numeric	10
Categorical	2
Boolean	1

Imagen 1. Overview del perfilamiento de los datos.

Con este reporte, observamos que el conjunto de datos cumple con los principios de completitud y unicidad, ya que no tiene datos faltantes ni filas duplicadas. Además, no se han detectado alertas de posibles errores en el formato de los datos de cada variable. Por lo tanto, no fue necesario realizar procedimientos adicionales relacionados con estos aspectos.

A continuación, se procedió a tratar las variables categóricas "Holiday" y "Functioning Day", cada una de las cuales tiene dos categorías. Debido a que estas variables representan opciones binarias ("No Holiday" o "Holiday", "True or False"), se codificaron como variables binarias (0 y 1). Por otro lado, la variable "Seasons" tiene cuatro categorías sin un orden significativo

(“spring”, “summer”, “autum” y “winter”). Para manejar esta variable, se decidió aplicar la técnica de one-hot encoding (dummies), creando una columna binaria para cada temporada. Se tomó “Autum” como la categoría base de referencia para evitar problemas de multicolinealidad en futuros análisis y modelos.

Holiday	Functioning Day	Seasons_Spring	Seasons_Summer	Seasons_Winter
0	1	0	0	1
0	1	0	0	1

Imagen 2. Visualización del perfilamiento de los datos.

Finalmente, los datos con las limpiezas y modificaciones realizadas se guardaron en un nuevo archivo “SeoulBikeDataClean.csv” para que pudieran ser procesados en otro cuaderno de Jupyter, permitiendo así la creación de los modelos de aprendizaje correspondientes.

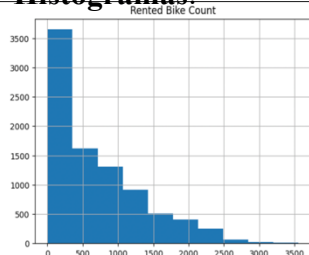
### Tarea 3 - Exploración de datos

En esta sección, se realizará un análisis exploratorio de los datos para describir estadísticamente y visualizar el comportamiento de las diferentes variables presentes en el conjunto de datos. El objetivo es identificar patrones, tendencias y relaciones iniciales que permitan extraer primeras conclusiones relevantes sobre el conjunto de datos.

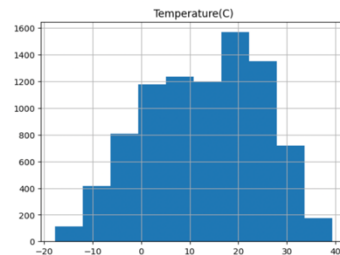
#### Estadísticas Descriptivas Clave:

- **Rented Bike Count:** Promedio de 704.6, máximo de 3556, desvío estándar de 644.9.
- **Temperature (C):** Promedio de 12.88°C, mínimo de -17.8°C, máximo de 39.4°C.
- **Humidity (%):** Promedio de 58.22%, variabilidad notable con un mínimo de 0% y un máximo de 98%.
- **Wind Speed (m/s):** Velocidad media de 1.72 m/s, con una variabilidad de hasta 7.4 m/s.
- **Solar Radiation (MJ/m2):** Radiación solar media de 0.57 MJ/m<sup>2</sup>, alcanzando un máximo de 3.52 MJ/m<sup>2</sup>.

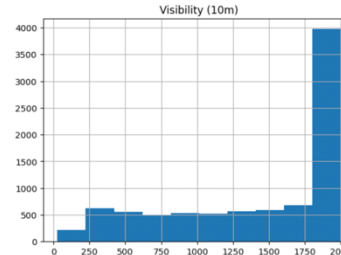
#### Histogramas:



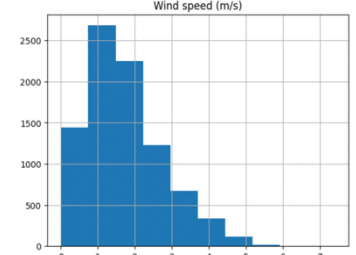
**Demanda Baja Predominante de Bicicletas Rentadas:** La mayoría de los días tienen un bajo número de bicicletas rentadas, con solo unos pocos días de alta demanda. Esto indica que la utilización del sistema de bicicletas es esporádica y puede depender de factores externos, como condiciones climáticas o eventos específicos.

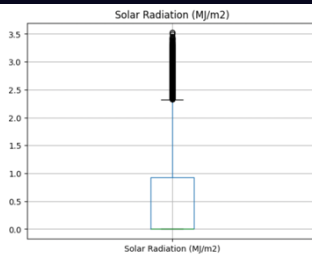


**Influencia significativa del clima sobre la demanda:** Variables como la temperatura, que sugieren que existen climas extremos en la región de estudio, probablemente pueden afectar negativamente la cantidad de bicicletas rentadas. Puesto que, situaciones climáticas extremas puede afectar la cantidad de gente que sale a las calles

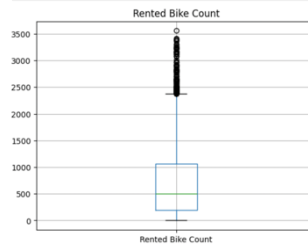


**Condiciones Predominantes de Alta Visibilidad y Baja Velocidad del Viento:** La alta visibilidad y baja velocidad del viento son las condiciones más comunes, lo cual es ideal para el ciclismo. Días con estas características pueden contribuir a un uso más consistente de bicicletas, mejorando la seguridad y comodidad de los usuarios.

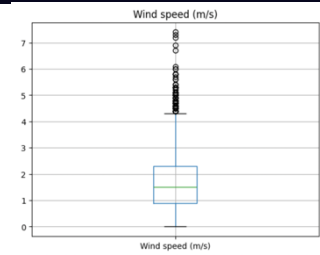




Similar al conteo de bicicletas rentadas, la radiación solar tiene una alta concentración de valores atípicos hacia el extremo superior. Esto indica que aunque la mayoría de los días tienen baja radiación solar, hay ciertos días con radiación extremadamente alta, lo que puede afectar la demanda de bicicletas por razones de confort y seguridad.



Gran cantidad de valores atípicos por encima del límite superior, lo que indica que existen numerosos días con un número de bicicletas rentadas significativamente mayor al promedio. Esto confirma la naturaleza esporádica de la alta demanda observada anteriormente y sugiere que ciertas condiciones o eventos pueden provocar picos en el uso.



La velocidad del viento presenta valores atípicos por encima del límite superior, lo que indica que días con vientos fuertes son excepcionales. Esto es importante, ya que puede tener implicaciones para la seguridad y comodidad de los ciclistas, afectando la demanda en esos días.

## Diagramas de dispersión:

- **Hora vs. Rented Bike Count:** Existe una relación positiva clara entre la hora del día y el número de bicicletas rentadas, alcanzando un pico en la tarde (alrededor de las 17:00-18:00), lo que indica una alta demanda posiblemente debido a los desplazamientos laborales.
- **Temperatura vs. Rented Bike Count:** La demanda de bicicletas aumenta con la temperatura hasta cierto punto (alrededor de 20-25°C), lo que sugiere que condiciones climáticas moderadas son ideales para el uso de bicicletas.
- **Snowfall vs. Rented Bike Count:** Hay una relación negativa entre la nieve y la cantidad de bicicletas rentadas, lo que indica que la presencia de nieve reduce significativamente la demanda de bicicletas, afectando la seguridad y comodidad del ciclista.

## Diagramas de violín:

- **Rented Bike Count:** La distribución muestra una asimetría a la izquierda con una alta densidad de datos en valores bajos de rentas de bicicletas. Esto sugiere que la mayoría de los conteos de bicicletas rentadas son bajos, con algunas observaciones que representan un conteo mucho mayor.
- **Temperature (C):** La distribución de la temperatura es aproximadamente simétrica alrededor de la media, con una mayor concentración de datos entre 10°C y 30°C. Esto indica que la mayoría de los registros de temperatura se encuentran dentro de este rango, siendo valores extremos menos comunes.
- **Wind Speed (m/s):** La distribución de la velocidad del viento tiene una alta concentración alrededor de valores bajos (cerca de 1-2 m/s) y disminuye rápidamente a medida que la velocidad del viento aumenta. Esto indica que, generalmente, las velocidades del viento son bajas, con pocas observaciones que superen los 5 m/s.

## Modelos

### Tarea 4 - Modelamiento

Después de limpiar, preparar y analizar los datos, se buscó un método para estimar el comportamiento de la variable dependiente (Rented bike count). En este caso, se desarrolló un modelo de regresión lineal que utiliza las 14 variables del conjunto de datos para estimar la cantidad de bicicletas rentadas por hora. Según Rumsey (2016), la regresión lineal es útil para entender y predecir el comportamiento de una variable dependiente con base en el comportamiento de una o más variables independientes. Posteriormente, se llevó a cabo un análisis estadístico para evaluar la efectividad del modelo y eliminar las variables que no contribuían significativamente a la explicación del comportamiento de la variable dependiente. Esto permitió obtener una estimación más precisa y ajustada a la realidad. A

continuación, se presentan los modelos generados y se detalla el modelo final seleccionado para interpretar la cantidad de bicicletas rentadas.

### Modelo inicial:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14}$$

Y: Rented Bike Count  
X1: Hour  
X2: Temperature(C)  
X3: Humidity(%)  
X4: Wind speed (m/s)  
X5: Visibility (10m)  
X6: Dew point temperature(C)  
X7: Solar Radiation (MJ/m2)  
X8: Rainfall(mm)  
X9: Snowfall (cm)  
X10: Holiday  
X11: Functioning Day  
X12: Seasons\_Spring  
X13: Seasons\_Summer  
X14: Seasons\_Winter

B0: 53.98370399970281  
B1 (Hour): 27.36755343612323  
B2 (Temperature(C)): 14.43826753216662  
B3 (Humidity(%)): -11.103888587537377  
B4 (Wind speed (m/s)): 19.496121345722678  
B5 (Visibility (10m)): 0.01745474269429792  
B6 (Dew point temperature(C)): 12.820752006513155  
B7 (Solar Radiation (MJ/m2)): -75.87692322007723  
B8 (Rainfall(mm)): -55.83998938115196  
B9 (Snowfall (cm)): 29.026841185932145  
B10 (Holiday): -115.79557975811028  
B11 (Functioning Day): 931.2135047213319  
B12 (Seasons\_Spring): -125.83055751023507  
B13 (Seasons\_Summer): -142.94948619323395  
B14 (Seasons\_Winter): -357.4897874909983

Para el modelo de regresión lineal se hace un entrenamiento con una división de los datos del 20-80. Esto con el fin de entrenar el modelo y después probar con los mismos datos que tan bueno es el ajuste para estimar el comportamiento de la variable dependiente. Esto es necesario para evitar sesgos de evaluación y obtener resultados más fiables y generalizables (Sangha, 2021)

Después se revisaron las medidas de desempeño del modelo para ver si las variables y el mismo modelo era significativo.

OLS Regression Results						
Dep. Variable:	Rented Bike Count	R-squared:	0.550			
Model:	OLS	Adj. R-squared:	0.550			
Method:	Least Squares	F-statistic:	764.8			
Date:	Wed, 04 Sep 2024	Prob (F-statistic):	0.00			
Time:	14:45:48	Log-Likelihood:	-65598.			
No. Observations:	8760	AIC:	1.312e+05			
Df Residuals:	8745	BIC:	1.313e+05			
Df Model:	14					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	35.5311	96.538	0.368	0.713	-153.706	224.768
Hour	27.5302	0.735	37.470	0.000	26.090	28.970
Temperature(C)	16.0705	3.662	4.388	0.000	8.892	23.249
Humidity(%)	-10.8090	1.030	-10.498	0.000	-12.827	-8.791
Wind speed (m/s)	19.2042	5.095	3.769	0.000	9.216	29.192
Visibility (10m)	0.0103	0.010	1.042	0.298	-0.009	0.030
Dew point temperature(C)	11.1647	3.835	2.911	0.004	3.647	18.682
Solar Radiation (MJ/m2)	-77.3099	7.593	-10.182	0.000	-92.194	-62.426
Rainfall(mm)	-58.4803	4.270	-13.694	0.000	-66.851	-50.109
Snowfall (cm)	32.6945	11.205	2.918	0.004	10.730	54.659
Holiday	-117.5786	21.605	-5.442	0.000	-159.929	-75.228
Functioning Day	932.1416	26.652	34.974	0.000	879.897	984.386
Seasons_Spring	-135.3105	13.880	-9.749	0.000	-162.518	-108.103
Seasons_Summer	-154.4829	17.211	-8.976	0.000	-188.220	-120.746
Seasons_Winter	-366.0543	19.708	-18.574	0.000	-404.687	-327.421
Omnibus:	1423.492	Durbin-Watson:	0.509			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2844.040			
Skew:	0.990	Prob(JB):	0.00			
Kurtosis:	4.968	Cond. No.	3.27e+04			

MAE: 326.4871946346789  
MSE: 190609.26508288088  
RMSE: 436.58820996779207  
R2: 0.5445663619093557

De los resultados anteriores, se determina que el modelo es globalmente significativo. En el caso de las variables, se concluyó que la visibilidad no es significativa. Por lo que, se realizó un nuevo modelo de regresión lineal excluyendo esta variable.

### Modelo final:

MAE: 326.60015025036734  
MSE: 190408.074931489  
RMSE: 436.357737334276  
R2: 0.5450470770653402

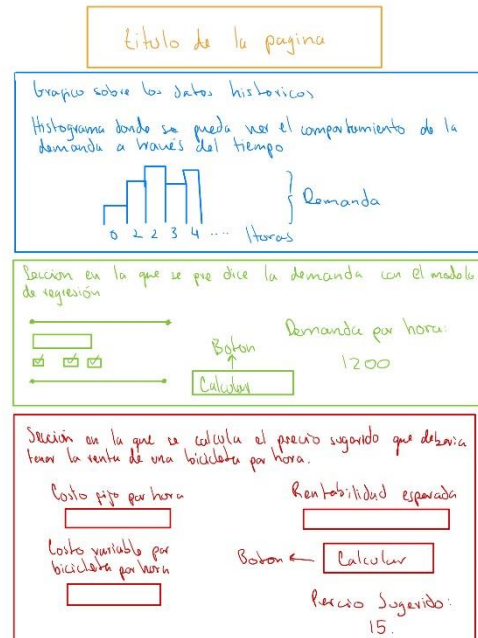
Los resultados de este modelo nos dan indicios que este permite estimar mejor la variable dependiente. Puesto que, el MSE dio un valor ligeramente menor. Por lo cual el nuevo modelo (sin la variable visibilidad) es mejor a la inicial.

Posteriormente, se analizó el VIF para determinar si posiblemente había problemas de multicolinealidad entre las variables independientes. De los resultados se pudieron sacar algunas conclusiones, pero estadísticamente estas al ser significativas no se excluyeron del modelo.

## Producto

### Tarea 5 - Diseño y desarrollo del tablero

Para esta tarea primero se realizó el wireframe a mano para poder tener una idea general de cómo iba a quedar el tablero. Acá se tomó la decisión de pedir por parámetro información financiera como costos fijos, variables y rentabilidad esperada para así poder dar un estimado del precio que debería tener la bicicleta en una hora.



En el producto final se le ofrece a la empresa una vista general de como se ha comportado la demanda a lo largo del tiempo dependiendo de la temporada. Se escogió la variable temporada ya que esta determina directa o indirectamente las otras variables como la temperatura, la velocidad del viento etc. Después se hizo un apartado en el que la empresa puede configurar los valores de las variables que están en el modelo para así aplicar el modelo de regresión lineal sobre estas condiciones para así poder otorgar un intervalo de confianza de 85% en el que se encuentra la demanda y la media de este intervalo la cual sería la demanda aproximada. Finalmente, la empresa puede ingresar por parámetro, como se dijo anteriormente, los costos fijos por hora, los costos variables por hora y por bicicleta y la rentabilidad esperada; con esta información la herramienta calcula un precio sugerido de renta de bicicletas en esa hora con las condiciones (configuración de las variables) preestablecidas. Este precio, con la demanda estimada, cubriría todos los costos y además dejaría la rentabilidad deseada. La fórmula que se utilizó para encontrar el precio es la siguiente:

$$\text{precio} = \frac{\text{Costo Fijo} + \text{Costo Variable} * \text{Demanda}}{\text{Demanda}} + \text{Rentabilidad Esperada}$$

## Tarea 6 - Despliegue

Para esta tarea, se utilizó el acceso a máquinas virtuales (MV) en EC2 a través de los recursos brindados por el laboratorio de aprendizaje de AWS academy para desplegar el tablero de datos. El tablero DASH fue implementado en una instancia de EC2, asegurando que fuera accesible y permaneciera en ejecución para los usuarios. Esta configuración permite que los interesados puedan visualizar los resultados de los análisis de datos de manera efectiva.



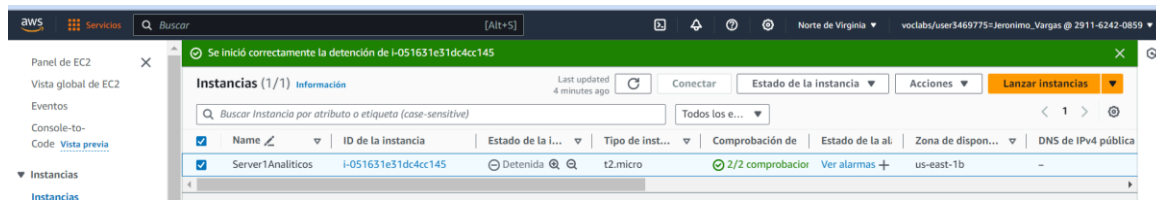


Imagen 233. Visualización de la creación de la instancia en EC2 para poder desplegar el producto.

## Reporte trabajo en equipo

En este proyecto, trabajamos de manera colaborativa para asegurarnos de que todos los miembros del equipo comprendieran las tareas que se estaban realizando en cada etapa del proceso. Aunque todos contribuimos al entendimiento general del proyecto, cada miembro asumió un rol de liderazgo en las áreas específicas seguridad de acuerdo con sus habilidades y conocimientos. Esta dinámica nos permitió ser más eficientes y asegurarnos de que cada tarea fuera manejada por la persona más adecuada.

**Jeronimo Vargas Rendon:** En el proyecto, mis principales aportes estuvieron en la ingeniería de datos y el despliegue del tablero en la nube. Me encargué de preparar, limpiar y transformar los datos para asegurar que estuvieran en un formato adecuado para el análisis y modelado, gestionando datos faltantes y tratando variables categóricas. También lideré el despliegue del tablero de datos en una instancia de EC2 en AWS, garantizando su accesibilidad y funcionamiento continuo. Además, colaboré estrechamente con mis compañeros para entender los procesos y contribuir con soluciones técnicas a lo largo de todo el proyecto.

**Carlos Gómez:** En el proyecto, mi aporte principal estuvo enfocado en realizar el modelo de regresión lineal adecuado para estimar la variable independiente, que en este caso son las bicicletas rentadas. Asimismo, busqué formas para mejorar el modelo realizado, ya sea buscando variables significativas y no significativas, minimizando el MSE o calculando medidas de desempeño. Esto con el fin de poder calcular el número de bicicletas rentadas de la mejor manera posible y más acercada a la realidad.

**Juan Esteban López:** En el proyecto, mis principales aportes estuvieron en la parte de análisis de negocio y el desarrollo del tablero. Con mis compañeros analizamos la situación planeada y decidimos solucionar el segundo enfoque propuesto que tiene un mayor componente financiero. Luego de decidir realice las preguntas que responderíamos a lo largo del proyecto. En cuanto al tablero, mis compañeros me entregaron un set de datos limpios y un modelo de regresión lineal el cual fue utilizado para mostrar gráficamente los resultados del proyecto y dar una visualización de las respuestas a las preguntas planteadas anteriormente en la tarea 1. En el tablero se puede destacar la predicción de la demanda que puede hacer el usuario al configurar el valor de las variables y el cálculo de un precio sugerido por bicicleta por hora para cubrir todos los costos del sistema y ser rentable.

## Referencias:

Sangha, K. (2021). *Why do we split the data into train and test sets?* Retrieved from <https://karansangha.com/posts/train-vs-test-split/>

Rumsey, D. J. (2016). *Statistics for Dummies* (2nd ed.). Wiley Publishing, Inc.

Repositorio GitHub: <https://github.com/Jeronimo2122/ModeloProy1.git>